specialbulletin från

PEDAGOGISK-
PSYKOLOGISKA
INSTITUTIONEN

LÄRARHÖGSKOLAN
MALMÖ

testkonstruktion och testdata

Bierschenk, I.:

FÖRSÖK MED AUTOMATISK SEPARERING AV REFERENSER I EN FLERSPRÅKIG DATABAS

Nr 34 November 1978

FÖRSÖK MED AUTOMATISK SEPARERING AV REFERENSER I EN FLERSPRÅKIG DATABAS

Inger Bierschenk

Bierschenk, I. Försök med automatisk separering av referenser i en flerspråkig databas. Testkonstruktion och testdata (Malmö: Pedagogisk-psykologiska institutionen), Nr 34, November 1978.

Denna arbetsrapport inom Informations- och dokumentationsprojektet (I&D) behandlar den fas i upprättandet av dokumentbaser som avser en automatisk separering av de olika språk som dokumenten i undersökningen är avfattade på. Det prövas och diskuteras på vilket sätt språkliga tecken i ett dokuments bibliografiska data bäst tas tillvara för utvecklingen av sökrutiner för automatisk separering. Det visar sig bl a att ett dokuments titel innehåller den lämpligaste informationen för detta ändamål samt att det engelska språket är lättast att automatiskt identifiera. Dessutom har kontroller av sökprogrammet gjorts, som bl a visar att en automatisk språkseparering tycks vara en väg att undvika vissa typer av fel som skulle uppstå vid en manuell kodning.

Nyckelord: Datalingvistik, datorbaserad lexikologi, forskningsinformation, I&D-system, informationslagring, informationsåtervinning.

INNEHÅLL Sid

1. PROBLEMFÖRANKRING OCH MATERIAL

Nya informations- och dokumentationssystem som bygger på användning av datateknik har installerats runt om i världen, men med mycket varierande effekter. Flera olika informationstyper har blivit tillgängliga jämfört med vad som vanligen kan hittas på bibliotek. Samtidigt har informationen för allt fler blivit mera svåråtkomlig. Ett av skälen är att systemen bygger på avancerade tekniska systemlösningar. Ett annat skäl är att det ännu inte har kunnat utvecklas program som kan hantera information som presenterats på många olika språk.

Projektet "Information och dokumentation (I&D)" vid pedagogisk-psykologiska institutionen har som ett av sina mål att utveckla program som tar hänsyn till strukturella förhållanden inom olika språk. Detta kräver att vi kan utveckla ett regelsystem som hjälper oss att separera dokumenten med avseende på sina språkområden. Nyckeln till detta är dokumentets bibliografiska beskrivning.

Denna rapport presenterar de enskilda stegen i processen att utveckla en algoritm. Syftet är att kunna åstadkomma separeringen automatiskt.

De dokument som utgör vårt basmaterial är 40 beteendevetenskapliga forskares egna verk och därtill kopplade referenser (i referenslistor och noter). De 40 forskarna utgör basmaterialet för ett tidigare projekt, där de intervjuades (se B. Bierschenk, 1974).

För att få en enhetlig form att arbeta efter skapades regler för på vilket sätt alla dessa referenser skulle läggas upp på magnetband. För den här framställningen är den bibliografiska representationen av intresse. Dokumenten (referenserna) redigerades enligt internationella konventioner (B. Bierschenk, 1973). Vilka problem som uppstod där har refererats i tidigare arbetsrapporter från projektet och behöver inte upprepas för förståelsen av denna rapport.

Hur en referens uppdelades för datamaskinell bearbetning visas i ruta 1.

Ruta 1. Representation av bibliografiska uppgifter för datamaskinell bearbetning

Post	Information om
1	Författarnamn med initialer för förnamn och uppgift om författarfunktion, t ex Ed
2	Titel och undertitel till ett dokument
3	Utgivningsort (tryckort)
4	Förlag
5	Utgivningsår, volym, häfte, sidangivelse
6	Tidskriftsnamn, rapportserie, stencil
7	Övriga dokumentkarakteristika

Posterna 2-6 är de intressanta för den här studien. Det är nämligen informationen inom dessa som är nödvändiga som identifierare av ett dokument och som vanligen anges i en referenslista.

De 40 forskarnas verk har fått entydiga identifieringskoder, liksom samtliga referenser som tillhör ett visst verk. Verkens antal är ca 800. Antalet referenser är omkring 19 000. (En kompletterande datainsamling startade i juni 1978.)

För den här presentationen är det viktigt att nämna att det vid kodningen tillämpades en princip med ett refereringsnummer istället för att hela referensen stansades upp, när kodaren visste att den förekommit tidigare i materialet. Referensen letades då upp i den löpande datautskriften och identifikationsnumret sattes in som referens. Meningen var att arbetet skulle gå snabbare och att identiska referenser skulle läggas till maskinellt i efterhand. För vår del betyder det nu att vi inte har vetat det exakta antalet referenser vid utprövningar och beräkningar. Det betyder också att inte alla 19 000 referenserna är olika referenser.

2. EXPLORATIVA UTPRÖVNINGAR

2.1 Språkbestämmande teckensträngar i specifika poster

För en första utprövning gjordes ett antagande om vilka ord eller teckensekvenser som kunde vara av betydelse för att bestämma en referens tillhörighet till ett visst språk. (En teckensträng - "string of characters" - kan definieras som en uppsättning element, vars enskilda delar används för att representera data.) Eftersom varje referens innefattar mellan 1 och 6 specificerade poster, där innehållet inom varje post tillhör en viss kategori inom en referens (se ruta 1) bestämdes att antaganden kunde göras i förhållande till varje posts innehållskategorier. (Post - "character subset" - kan definieras som ett urval från en uppsättning av tecken, som används för ett visst syfte.) Det betydde i denna första utprövning att vi kunde arbeta med kortidentifieringen som utgångspunkt. Det antogs då att post 3 (tryckort) eller post 6 (tidskrift, institutionsserie o likn) skulle ge önskat utfall. Vi antog därmed att tryckorten Stockholm anger en svenskspråkig referens och t ex New York en engelskspråkig. Likaså kunde ortsangivelse inom post 6 ange svensk institutionsrapport, som är vanligt förekommande i materialet. Post 6 har också vissa typiska drag, som t ex rapport, tidskrift, tidning respektive Education, University, Psychology.

Vi försökte täcka in så många språk som möjligt av dem som vi visste förekommer i materialet. Därför gjordes liknande antaganden angående tyska, franska, finska, danska, norska och italienska. Från början visade det sig att de svenska ortsangivelserna inom post 6 slog fel ut. Sökningen gav engelska skrifter utgivna på svenska institutioner. Vetskapen om detta gjorde att vi plockade bort de svenska universitetsortnamnen från sökorden för post 6. De strängar som bildade våra första sökord visas i ruta 2. Asteriskerna betyder en s k trunkering, där man garderar för att strängen är del av ord, dvs det kan förekomma tecken både före och efter.

I instruktionerna sades också att när ingen asterisk förekommer betyder det att blanktecken följer. Står ingen asterisk framför föregås strängen av blank eller utgör textens början, dvs kolumn 21. Asterisk efter kunde också betyda att punkt följer.

För att få en uppfattning om utfallet bestämdes att frekvenser skulle beräknas på varje söksträng.

För tydlighetens skull poängteras att orden "sökord" eller "söksträng" används för att beteckna de sekvenser av tecken som programmet söker med för att diskriminera mellan språken.

Ruta 2. Teckensträngar och sökord för språkseparering: Allmänna explorationer*

1. Antagen svenskspråkig referens:

inom post 3 återfinns:	inom post 6 återfinns:	
Stockholm:	ped*	Dagens Nyheter
Lund:	psyk*	korr*
Malmö:	*skrift*	arbet *
Uppsala:	*tidning*	Sydsvenska Dagbladet
Göteborg:	SOU	
Linköping:	*skol*	
Norrköping:	*rapport*	
Umeå:	tidskr*	
Helsingfors:	*förening*	
Örebro:	social*	
Falun:	opubl*	

2. Antagen engelskspråkig referens:

inom post 3 återfinns:		inom post 6 återfinns:	
New York:		Education*	
Chicago:		Educ. *	
London:		Psychology	
Washington:		J	2)
Pittsburgh:		Journ.	2)
Baltimore:		Journal	2)
Los Angeles:		Bull*	
Boston:		Rev. *	
San Francisco:		Review	
Evanstone:		Diss*	
Glencoe:		Mimeo*	
Paris:	1)	University	

inom post 4 återfinns:

University
Press
Stanford
Harvard

1) Under förutsättning att post 4 eller 6 innehåller UNESCO
2) Under förutsättning att efterföljande sträng (ord) inte är de

* Innebär att flerspråkighet förekommer (jfr även kriteriedisk., s 14)

Ruta 2. (forts)

3. Antagen tyskspråkig referens:

inom post 3 återfinns:	inom post 6 återfinns:
Leipzig:	z tschr*
Berlin:	z
Heidelberg:	z scht
Zürich:	F
München:	F. *
Jena:	ü*
Bern:	Pädag*
Haag:	
Halle:	
Frankfurt:	
Weinheim:	
Tübingen:	
Dortmund:	
Breisgau:	
Freiburg:	
Hamburg: 3)	

3) Under förutsättning att post 4 eller 6 inte innehåller UNESCO

4. Antagen franskspråkig referens:

inom post 3 återfinns:	inom post 6 återfinns:
Paris: 4)	de
	des

4) Under förutsättning att post 4 eller 6 inte innehåller UNESCO

5. Antagen finskspråkig referens:

inom post 3 återfinns:

Helsinki:
Jyväskylä:

6. Antagen danskspråkig referens:

inom post 3 återfinns:	inom post 6 återfinns:
Köbenhavn:	dansk
Köpenhamn:	

7. Antagen norskspråkig referens:

inom post 3 återfinns:	inom post 6 återfinns:
Oslo:	norsk

8. Antagen italienskspråkig referens:

inom post 3 återfinns:

Roma:

Det ska också sägas att ambitionen inte var att vara heltäckande. Det skulle ha betytt att vi listat t ex samtliga ortskoder innan specificeringen gjordes. Sökorden exemplifierar några vanliga förekommande ord eller teckensekvenser ur det delmaterial som studerats samt baserar sig på kännedom om materialets utseende och kodning.

2.2 Första utprövningen

Utprövningen skedde på den s k verkbasen, som innefattar våra 40 forskares egna skrifter. Ett typiskt drag är bl a att den största produktionen föreligger i rapportform, vilket medför att post 6 haft stor betydelse.

Sökrutinen har varit att det första ordet som påträffats inom respektive post har varit bestämmande för språksepareringen. Posterna har avsökts i ordningsföljd, dvs om post 3 inte givit utslag ("träff") i en referens har sökningen fortsatt i de angivna posterna. De förbehåll som anges för vissa sökord i ruta 2 (t ex Paris) har inte tagits hänsyn till. Det hade komplicerat programmen, vilket bedömdes vara onödigt i det här skedet.

För att kontrollera utfallet lät vi skriva ut verken såsom de inlagts i verkbasen. Efter varje referens (och varje post) angavs i en tvåsiffrig kod dels vilken språkkod som tilldelats referensen, dels vilken post som givit utfallet. Därmed kunde kontrolleras såväl korrekta som felaktiga utfall. Exempel på utskrift ges i ruta 3. (Författarnamnen är uteslutna.)

Det korrekta exemplet visar svensk språkkod (1) och att utfallet följt av ped* inom post 6. Det första felexemplet har registrerat F. i post 6 som skulle betyda tysk referens (se ruta 2). Detta F. betyder här "för", vilket visar att en språkblandning blir resultatet. En dator kan ju i det här fallet inte "förstå" skillnaden. Det andra exemplet har utfallit via ortskoden i post 3 och har oriktigt registrerats som svenskt. Det tredje har inte fått någon språkkod, eftersom ingen sträng av de i ruta 2 angivna har påträffats. Exemplet visar dels att post 3 inte förekommer samtidigt med post 6 (generellt), dels att vissa tidskriftsnamn inte bildas på det vanliga sättet, dvs för tyska språket i form av "Z(eitschrift) f(ür)..." eller liknande.

Ruta 3. Exempel på utskrift från verkbasen

Post ident	Referensinnehåll	Språk best
Korrekt utfall:		
211	Är evaluering forskning?	16
511	1971, 0107, 0002.	16
611	Pedagogisk Tidskrift.	16
Felutfall:		
221	Standardprovfrågan med särskild anknytning	46
222	till tyska och engelska från sjunde skolåret.	46
511	1960.	46
611	Riksfören. f. lärarna i moderna språk.	46
122	Swedish Council for Social Science Research.	13
221	Two decades of educational research.	13
222	Social Science Research in Sweden	13
311	Stockholm:	13
411	The Swedish Council for Social Science Research,	13
511	1972.	13
221	Probleme der Schuldemokratie: Forschungsaufgaben	
222	und Schwedische Ausgangspunkte.	
511	1968. 0005	
611	Didakometrie und Soziometrie.	

Språkbestämningskoder: Första kolumnen anger vilket språk det gäller; 1 = sv, 2 = no, 3 = da, 4 = ty, 5 = eng, 6 = fr, 7 = it, 8 = fi. Andra kolumnen anger vilken post som bidragit till utfallet. (Utskriften är ej identisk med datorutskriften, som t ex inte har gemena bokstavstecken.)

Några beräkningar har inte utförts på dessa explorationer. Istället gicks verkbasen igenom (ca 470 verk). Det gällde nu att på basis av dessa utfall ställa några hypoteser angående vilka teckensekvenser som var de lämpligaste att använda för att språkligt dela upp våra baser. En första åtgärd var att använda samma sökprocedur på referensbasen (med ett urval av samma storlek som verkbasen) för att få en uppfattning om eventuella skillnader. Dessutom skulle vi kunna få så många alternativa val som möjligt. Ett prov gjordes i samband med denna bas: vi antog att det lilla ordet <u>of</u> hade stor verkan vid separeringen på engelska.

Istället för Journal och flera andra ord (Journal kan också vara franska) kunde i första hand of användas i post 6.

Referensbasen har gåtts igenom mera systematiskt med avseende på de olika ordens utfall. Vi ska här inte gå in i detalj utan endast ange de mest väsentliga resultaten (jfr sökorden i ruta 2). De baserar sig på automatisk frekvensräkning.

I den svenska basen har ped* , skol* och rapport haft den ojämförligt största effekten för utfallet inom post 6, därefter följer psyk* . Det ger en bra bild av ämnesområdets institutionella förankring. I fråga om tryckorten (för böcker) dominerar Stockholm starkt. Andra övriga orter tycks inte ha en tiondels andel av vad Stockholm har.

Norska och danska utfall är få. Endast Oslo verkar ge något väsentligt. Tyska referenser har knappast förekommit i provmaterialet. En "träff" på Z har registrerats. Franska, italienska och finska har inte förekommit, enligt beräkningarna.

Det största antalet totalt sett finns i den engelska basen. Lika dominerande som Stockholm tycks vara för svenskt språk är New York för engelskspråkiga referenser. Övriga orter har mycket liten betydelse förutom London. När det gäller post 6 har vårt prov med of slagit bäst ut. Det får dubbelt så många "träffar" som närmast följande Review. Därefter följer Educ* , University och J. Dessutom har post 4 någon liten betydelse när det gäller amerikanska universitetsförlag av typen Harvard University Press. När post 3 inte har gett utfall, så kan alltså närmaste post (4) hjälpa till.

Felaktiga utfall syns emellertid inte i denna frekvensräkning (se typer i ruta 3).

Eftersom det visade sig att of hade god effekt, kunde det antas att det även på andra kortnummer hade samma verkan. Post 2 kommer då i förgrunden. Dessutom såg vi (ruta 3) att tryckorten hade många felutfall. Doktorsavhandlingar och andra boktryck ges t ex ut i Stockholm på engelska och många liknande fel skapar onödigt brus i baserna. Den manuella genomgången och omkodningen skulle ta för mycket tid i anspråk om sådana tydliga felkällor inte togs bort.

En beräkning gjordes på (1) hur många korrekta referenser som listats, (2) hur många felaktiga utfall som inträffat och (3) hur många

som återstår, kallade "övriga". De övriga är således sådana som p g a programmets begränsningar inte automatiskt har bestämts. De ska inte förväxlas med de felaktiga. Det exakta antalet referenser i provmaterialet är 733. Proportionerna visas i tabell 1.

Tabell 1. Språkbestämning av referenser: Utfall av sökning inom posterna 3, 4 och 6. Provmaterial

	f	%
Korrekta utfall	572	78.03
Felaktiga utfall	27	3.08
Övriga	144	19.73
Σ	733	

Som tabellen visar är felen inte många. Men andelen referenser som inte kodats är tillräckligt många för att vi ska behöva justera i sökorden. Frågan uppstod på vilket sätt denna justering ska ske. Blir utfallet bättre om vi lägger till ett antal söksträngar, t ex flera tryckorter för att post 3 ska ge mera? Eller bör vi försöka vara entydigare i sådana karakteristika som är typiska för ett visst språk, bortsett från vad som är typiskt för referenser? Användningen av <u>of</u> tydde på att det kunde vara en väg. Samtliga post 2 studerades därefter, eftersom det är titeln som ger "cues" till entydiga språkdrag. Språkdrag kan då definieras som strukturella drag, dvs sådana tecken som tillhör den syntaktiska uppbyggnaden. Vi kan ta ett exempel från titeln "On learning and human ability", där prepositionen <u>on</u> och konjunktionen <u>and</u> tillhör sättet att konstruera meningar. De andra orden tillhör språkets innehållsliga enheter. Dessa har emellertid också syntaktiska drag, t ex att <u>learn</u> bildar verbalsubstantiv genom suffixet <u>-ing</u>. För att kunna instruera en dator att tilldela denna referens till en engelsk bas fordras att datorn har tillgång till t ex engelska prepositioner eller engelska suffix, en lista (lexikon) som inte skulle bli särskilt stor. Om vi skulle vilja att ordet <u>ability</u> utgjorde "cue" till en

engelsk referensbas, fordras däremot att datorn har i sitt minne samtliga engelska substantiv (i betydelsen uppslagsord, se Allén, m fl, 1977, (kap 4). Visserligen är det möjligt att bestämma denna vokabulärs utseende och omfång, men det vore opraktiskt att låta minst åtta språks lexikon utgöra sökorden för en matchning mot det empiriska materialet. Dessutom blir söktiden längre med strängarnas längd. Alltså kan man tänka sig att vissa "småord" eller "formord" är en bättre väg. Dessas antal är också begränsat i varje språk (förnyelse sker knappast). Vi måste bara försöka undvika att ett ord inte också finns i flera språks lexikon, t ex de vanliga <u>den</u>, <u>in</u> etc. För att illustrera problemet visar ruta 4 en matris med exempel på formord som kan förekomma som teckensekvens i fler än ett språk. Exemplen är hämtade ur provmaterialet.

<u>Ruta 4.</u> Exempel på formord som kan förekomma som teckensekvens i flera språk

Ord (tecken-sekvens)	Språk							
	sv	da	no	eng	ty	fr	it	fi
om	x	x	x					
for	x	x	x	x				
in	x			x	x		x	
to		x	x	x				
på	x	x	x					
mot	x		x					
i	x	x	x					
an	x			x	x			
med	x		x					
av	x		x					
a				x		x	x	
under	x	x	x	x				
at		x	x	x				
att	x		x					
vid	x	x	x					
den	x	x	x		x			
der		x	x		x			
la						x	x	
des					x	x		
ja	x	x	x		x			x
ne						x	x	x

Utan att gå in på olika betydelse och funktion hos orden i olika språk kan vi tydligt se att en hel del ord, betraktade som en kombination av tecken, kan förekomma i flera språk. Noggrannare ordbokskontroller än i våra exempel skulle kanske ge flera likheter än vad vår matris exemplifierar. Datorn "vet" emellertid ingenting om dessa likheter utan kan bara reagera på våra instruktioner. Om vi alltså ger instruktionen att sortera in en titel i den svenska basen om i titeln påträffas strängen in, så är det högst sannolikt att denna bas till största delen kommer att innehålla tyska och engelska titlar. I båda språken finns in i likartad funktion.

Vi kan också ha andra fall, där teckensekvensen är en preposition i ett språk och t ex ett substantiv eller ett verb i ett annat (t ex mot, som på norska betyder "mod", for som på svenska är preteritum av verbet "fara").

De olika funktionerna är olika vanliga rent generellt i språket, vilket vi bör ta hänsyn till vid en bedömning av rimliga konsekvenser av enskilda sökord. Det är ju bl a därför som prepositionernas strukturella egenskaper är den främsta utgångspunkten i det här försöket. Vår kännedom om ämnesområdet kan också hjälpa till att utforma en sökprofil med ett så entydigt utfall per specificerat språk som möjligt. För diskussionen kan uppställningen i ruta 5 vara till hjälp:

Ruta 5. Ord som tecken: Dimensioner i en flerspråkig databas

	1	2
Specifik	ett språk en funktion	ett språk fler än en funktion
	3	4
Ospecifik	fler än ett språk en funktion	fler än ett språk fler än en funktion

De två första dimensionerna torde vara oproblematiska vid språkspecifik sökning. Om ett ord (teckensekvens) har flera funktioner men begränsade till endast ett språk, så fungerar det ändå som sökord. Ord som kan hänföras till dimensionerna 1 och 2 kan tas med i ett separeringslexikon.

Dimensionerna 3 och 4 är båda problematiska för vårt vidkommande (se ruta 4). Båda är för ändamålet ospecifika. Ju fler språk och funktioner som är inblandade, desto sämre är sekvensen. Det visar sig emellertid att inte enbart antal språk behöver vara avgörande vid beslut om vilka ord som kan ingå i ett separeringslexikon. Andra faktorer kan vara med och avväga. Ett ord med samma funktion i flera språk kunde t ex vara högt frekvent i ett av dem men till synes nollfrekvent i andra (räknat på ett testmaterial). Ett annat exempel är att förekomsten endast avser två språk, där funktionen är olika i vart och ett av språken. I det första fallet gäller det att avgöra om potentiell förekomst räcker för att utesluta ordet. I det andra vägs innehållsliga faktorer in. Några kriterier för våra separeringsförsök ställdes därför upp.

Ett ord togs med i ett separeringslexikon

1. om det väntades förekomma i bara ett språk, oavsett om det kunde ha fler än en funktion,
2. om det skulle kunna förekomma i ett högfrekvent språk och samtidigt i ett lågfrekvent språk, men i annan funktion,
3. om det skulle kunna förekomma i två högfrekventa språk, men med en innehållslig funktion i det ena språket, som inte förväntades förekomma,
4. om det skulle kunna förekomma i både högfrekvent och lågfrekvent språk i samma funktion, men med så proportionellt hög frekvens i det högfrekventa språket att det antogs innebära större förlust för utfallet att inte ha ordet med än att titlar från ett lågfrekvent språk blandades in.

Exempel på kriterium (1) är <u>of</u>, på (2) <u>att</u> (norska "åter"), på (3) <u>and</u> (fågeln "and" på svenska) och på (4) <u>hos</u> (jfr övr nordiska).

Exempel på överväganden där ett ord inte togs med är <u>mot</u>. Det förekommer i ett högfrekvent och samtidigt ett lågfrekvent språk i materialet (vilket våra preliminära beräkningar tyder på). Detta ord betyder på norska "mod". Vi kan inte utesluta att litteraturen från Norge handlar om mod. Referenser i beteendevetenskap kan handla om en hel del, som vi inte omedelbart förknippar med området. Mod är emellertid en psykisk företeelse, en abstraktion som inte kan uteslutas. Det kan däremot <u>and</u>, som är konkret och tillhör ett annat område (fåglar).

Läsaren görs här uppmärksam på att det finns möjligheter att arbeta på en ännu "högre" nivå, nämligen den grafotaktiska, som försöker att

med hjälp av vissa unika grafemkombinationer specificera olika språk (t ex distinktionen kk/ck). Denna nivå studeras inte här (se vidare kap 3. 3).

Såväl själva kriterierna som besluten bygger på antaganden som gjordes med hjälp av ett provmaterial. Några vanliga ordböcker och ordlistor användes för enstaka kontroller.

Nästa steg var att studera utfallet av en sökning i dessa 733 referenser inom post 2. Vi skulle sedan kunna jämföra resultatet av titelsökningen (språkspecifika "cues") med sökningen inom posterna 3, 4 och 6 (referensspecifika "cues").

För att få bättre överblick vid nästa utprövning togs bara de troligen största språken svenska och engelska ut i en första omgång. En lista ges i ruta 6. Till dessa ord lades ett par sökord, som torde vara centrala för många studier inom området och som bör kunna fungera när andra inte gör det.

Ruta 6. Språkspecifika ord för sökning av svenska och engelska titlar

Titeln är svensk om inom post 2 återfinns:		Titeln är engelsk om inom post 2 återfinns:	
att	mellan	and	some
eller	och	as	the
ett	samt	from	towards
från	som	how	who
för	till	is	with
hos	ur	its	
hur	vad	of	
inför	vem	on	
inom	är	or	
dessutom: skola*		school*	

Orden togs fram i genomgången av listan och har inte anspråk på att vara typiska annat än för provmaterialet (flera skulle ha kunnat listas ur respektive språk).

För att kunna bedöma rimligheten i att nästan uteslutande använda post 2 i sökningen gicks referenserna igenom, så att en anteckning gjordes för varje referens som skulle ha fallit ut genom sökning inom post 2 med orden ur ruta 6. Sådana antagna korrekta utfall kan då diskuteras med hänsyn till resultatet i tabell 1 (s 11). Eftersom en preli-

minär beräkning inte är direkt jämförbar ska endast några tendenser anges: Antalet korrekta utfall vid första utprövningen var 78 %. Sökningen företogs då med ett större antal sökord och innefattade posterna 3, 4 och 6. Post 6 hade därvid stor betydelse, dvs verkbasen innehåller många institutions- och tidskriftsartiklar. Om vi nu söker igen med några få sökord för engelska och svenska referenser inom titlarna med tillägg av New York och London för engelska böcker samt ordet Press för post 4, så tycks det som om omkring 80 % korrekta utfall skulle uppnås, och då har vi ändå inte post 6 med. Eftersom vissa tvetydigheter finns på post 6, t ex att Educ* också ger svenska rapporter, kan alla sådana "överblivna" referenser studeras separat.

Till instruktionerna för en förnyad utprövning lades önskemålet att separeringen skulle göras hierarkiskt, dvs de svenska referenserna skulle tas ut i en första omgång, sedan de engelska. Därmed undviks sådana (om än få) fall där ett engelskt begrepp, t ex namn på ett test, studeras i en svensk rapport eller artikel. Det första ordet i en titel kan därmed vara 'the' som skulle tillordna referensen till en engelsk bas. Tillägget visade sig inte möjligt att göra i det här skedet.

2.3 Andra utprövningen

En första kontrollkörning på ca 1 000 kort gjordes. Det visade sig att utfallen via post 2 dels har största andelen utfall, dels är helt korrekta (n = 147 av totalt 187 referenser). Tillsammans med några få korrekta utfall på post 3 (för engelska New York och London) utgör de 86 % på detta lilla material.

Nu gjordes en förnyad sökning, denna gång på ett antal av ca 8 000 kort för beräkningar av utfallen och för att en specificering ska kunna ske mot bakgrund av felen och de obestämda. Detta större material innehåller totalt 1 384 referenser, som fördelar sig så att antalet korrekta svenska och engelska utgör 1 101. 283 referenser har inte utfallit via sökorden. Korrekta utfall ser i detta material ut att utgöra ca 80 %, dvs omkring 6 % skillnad jämfört med ett material på 1 000 kort. Vad som främst tycks orsaka att felutfallen ökar något är dels att författarna skiljer sig åt, dvs lite längre fram i materialet visar det sig att ett par författare har en mer differentierad referenslista, dels att

en person (vi har kommit fram t o m person nr 4) har tyskspråkiga referenser, vilka inte alls har förekommit på de första 1 000. De tyska referenserna är flest av övriga språk hittills i materialet. Av felutfallen utgör de ca 9 %. För att reducera felen kan därför tyska sökord läggas till. Utsikten att en fortsatt sökning på ytterligare material (eller hela materialet) reducerar antalet obestämda är stor. Det beslöts att tyska söksträngar skulle läggas till och att dessa tillsammans med övriga sökord skulle testas på en del av materialet, där det förväntades variationsrikedom i referenserna, dvs person 09 och 20. Dessutom skulle vi lista språken för sig, engelska, svenska och tyska samt en extra lista för "övrigt" för den fortsatta överskådlighetens skull. De tyska tilläggen anges i ruta 7.

Ruta 7. Språkspecifika ord för sökning av tyska titlar

Titeln är tysk om inom post 2 återfinns:	
auf	und
das	von
die	über
für	zum
im	zur
mit	
dessutom: schul*	

En gardering för kort 6 gjordes mot bakgrund av de första testningarna, nämligen att förkortningen Z för tidskriftsnamn togs med. Däremot har inga tryckorter angivits på tyska. Dessa tycks vara många och ingen dominerar lika starkt som New York för de engelskspråkiga.

2.4 Tredje utprövningen

Andelen referenser för de sökta språken och restutfallet visas i tabell 2.

Tabell 2. Språkbestämning av referenser:
Utfall av sökning inom post 2. Delmaterial

	Eng f	Eng %	Sv f	Sv %	Ty f	Ty %	Övr f	Övr %	Σ f	Σ %
Korrekt	1617	38	824	19	345	8			2786	65
Fel	9	0	2	0	3	0				0
Övriga							1469	34		34
Σ	1626	38	826	19	348	8	1469		4269	99

Procenttalen har avrundats till närmaste heltal

De nio engelska felen har flera orsaker: sex fel beror på att J inom post 6 också samlat franska tidskrifter. Detta har tidigare konstaterats, men sökordet hade ej borttagits. Dessutom hade samma sökord givit en finsk referens. Kort 4 Press har givit mycket få utfall, däribland ett fel, som var en norsk referens. De fel som registrerats i den svenska listan är två som, som visat sig vara danska titlar, ett förbiseende vid specificeringen av sökorden. Det tyska sökordet von gav två svenska referenser, där på titelkortet finns ett personnamn med von. Von togs med enligt det första kriteriet. För säkerhets skull tas det bort igen. Någon författare kan ha skrivit bibliografier över en person, vars namn då förekommer i titeln. Det tredje felet beror på att en finsk titel har en tysk översättning tillagd. Post 4 utgår nu ur försöken.

Vi ser här att den s k restlistan (kategorin Övrigt) är förhållandevis stor. Det beror på att de två personernas referenser är mycket variationsrika, både till innehåll, form och språk. Det senare är mest betydelsefullt på det här stadiet. Mängden tyska referenser är stor jämfört med tidigare utprövningar. Dessutom finns mycket franska och latin.

Eftersom restlistan var relativt stor, bestämdes att den nya sökningen skulle ske enbart på den och samtidigt vara ett test innan hela materialet togs med. Med hänsyn till de fel som upptäcktes har sökorden från rutorna 5 och 6 justerats. Som skulle egentligen inte ha kvalificerat sig. Frekvensen på svenskt material (kriterium 4) var för låg.

De olika språkens andel har studerats, där det visar sig att svenska, engelska och tyska tycks uppta närmare 80 %. Latin har en förvånansvärt hög andel (6 %), vilket beror på att det förekommer ofta hos en enda författare. Men eftersom latin dels inte kan hänföras till särskilda länder (utsorterade via tryckorter), dels inte är spritt över flera författare, görs inga försök till sökord. Franska däremot kan finnas hos flera personer, likaså norska och danska. Övriga språk upptar bara 1 % (däribland finska och italienska, se ruta 2) och blir utan sökord.

Det är tydligt att vi behöver arbeta med ett mer ändamålsenligt synsätt för att få denna språkseparering färdig. (En fortsatt metodutveck-

ling kan ske sedan, när samtliga i projektet involverade har tillgodosetts för sitt fortsatta arbete.) Redan tabell 2 visade att felen inte blir många, vilket betyder att de kan tas ut manuellt och insorteras i efterhand i rätt språkbas.

Med det ändamålsenliga för ögonen innebär det också att vi som sökord på titel måste införa ämnesord (förutom skola* , etc) för att undvika en för stor restlista. Ämnesorden ger ensamma en mängd referenser främst av typen läroböcker och handböcker, alltså med korta titlar. Dessa ord är då typiska för materialet, t ex *psykologi* , Erziehung* , child* . *psykologi* finns i de tre skandinaviska språken och där kommer en blandning att ske. Det är emellertid en snabbare väg att låta dessa fel förekomma och rätta dem manuellt, än att just nu försöka prova ut på vilket sätt referenserna ändå kommer ut korrekt. Det visar sig alltså att många problem uppstår när man önskar att ett visst sökord ska ge fullständigt korrekta utfall. Sådant är svårt att uppnå när söklogiken är den enklast tänkbara (jfr ruta 2, och s 16 där inskränkningar, som skulle ha betytt programmeringssvårigheter, inte togs med). Om vi t ex hade velat att *psykologi* skulle fungera 100%-igt, hade det behövts en tilläggsregel som t ex säger att post 3 inte samtidigt får innehålla Oslo, eller att sökning skulle ske hierarkiskt.

Sådana pragmatiska ställningstaganden medförde utökat antal sökord och sex sökta språk.

2.5 Fjärde utprövningen

Innan vi lät programmet verka på hela referensbasen, testades det på restlistan från föregående utprövning. Några justeringar behövde göras i sökordens trunkeringar och några stansfel (stavfel) rättades. Därefter har hela referensbasen (dvs de fullständiga referenser som ligger i samma fil) genomsökts. Vi har fått ut en lista för vart och ett av språken svenska, norska, danska, tyska, engelska och franska. Alla andra språk samt av annat skäl oklassificerat finns på en särskild lista, som språkbestämts manuellt.

De sökord som har kommit till användning i denna sökning presenteras i ruta 8. Resultatdiskussion sker därefter.

Ruta 8. Teckensträngar och sökord för språkseparering av referenser

1. Svenskspråkig referens

inom post 2 återfinns:

är	till	års*	*svensk*
ur	inom	läro*	*arbet*
och	samt	barn*	*undersökning*
för	från	Sverige*	*utbildning*
att	inför	beteende*	*utredning*
hos	eller	uppfostr*	*mätning*
ett	mellan	begåvning*	*historia*
vad	betänkande	personlighet*	*psykologi*
vem	någ*	*skola*	*pedagogiska*
hur	vux*	*skolor*	*pedagogik*

inom post 6 återfinns:

SOU	prop*	*svensk*
och	arbet*	*utredning*
års*	lärar*	

2. Norskspråkig referens

inom post 2 återfinns:	inom post 3 återfinns:
norsk	Oslo
	Kristiania
	Krist.

inom post 6 återfinns:

Oslo
norsk

3. Danskspråkig referens

inom post 2 återfinns:	inom post 3 återfinns:
undersögelse	Köbenhavn
	Köpenhamn
	Odense

inom post 6 återfinns:

Kbhn
dansk

Ruta 8. (forts)

4. Tyskspråkig referens

inom post 2 återfinns:			inom post 3 återfinns:
im	mit	schrift*	Leipzig
und	über	bildung*	Berlin
das	oder	erziehung*	Frankf.
zur	ohne	pädagogisch*	
zum	nach	untersuchung*	
für	pädagogik	psychologisch*	
bei	geschichte	* jugend*	
als	ein*	*unterricht*	
vom	werk*	*buch*	
auf	schul*	*forschung*	
die	deutsch*	*wissenschaft*	

inom post 6 återfinns:

und
schrift

5. Engelskspråkig referens

inom post 2 återfinns:

of	how	studies	psychological
is	who	learning	child*
on	with	research	teach*
by	from	training	adult*
or	some	analysis	school*
as	into	education	measur*
the	study	psychology	reading*
and	towards	personality	america*
its	methods	educational	

inom post 6 återfinns	inom post 3 återfinns:	
Review	London	New York
	Chicago	Pittsburgh
	Oxford	San Francisco

6. Franskspråkig referens

inom post 2 återfinns:		inom post 3 återfinns:
ou	pour	Paris
un	chec	
les	France	inom post 6 återfinns:
sur	l'	
une	oeuvre	Paris
dans	française	Revue

De tyska orden, utom ortnamnen, har ej markerats med versaler, eftersom vår tekniska utrustning inte medger denna skillnad.

Efter våra första explorationer har vi alltså kommit fram till att dessa sökord ger oss ett tämligen bra resultat. Jämfört med ruta 2 (s 7) ser vi att post 3 och 6 har liten betydelse nu och att våra huvudsakliga ord finns inom titeln (post 2). Några sökord gav inget utfall (ca 15 st). Det gällde främst några specificerade på post 6, där tydligen referensen har klassificerats utan dessa ord, t ex 'Z' som lades till. De redovisas inte här. Som vi kan se har ord från post 6 blivit mycket få.

Resultatet från denna utprövning, som tills vidare får betraktas som slutgiltig, ska nu redovisas i detalj. Först presenterar vi lite siffermaterial och sedan vad dessa siffror står för. En överblick över fördelningen av hela utfallet ges i tabell 3 (s 23).

Kategorin Övriga anger de referenser som inte blivit bestämda genom sökprogrammet, dvs "cues" har saknats. Vi kan jämföra med provmaterialet från tabell 1. Där ser vi att andelen korrekta nu stigit med närmare 10 %, troligen till följd av att sökning i titelposten är säkrare än i de övriga. Tabellen ger dessutom fördelningen korrekta och felaktiga utfall för de sex språken.

Av tabellsiffrorna kan vi bl a utläsa att referenser på engelska och svenska är mest förekommande och att de dessutom har minst antal fel. Vi ser också att franska tycks ha ovanligt många felutfall. Vad felen har berott på redovisas i ruta 9 (s 24).

Av felanalysen i ruta 9 ser vi att de flesta fel finns bland de svenska med *psykologi* och *pedagogik* som sökord. Det var också väntat och dessa fel har vi tagit hand om manuellt. Andra fel i den svenska basen var inte lika väntade. Några ska kommenteras: Att och förekommer i norska och danska var inte väntat. Flera ordlistor som använts tar inte upp ordet, och när det tas upp sägs det att ordet på norska eller danska stavas og. Däremot har hos funnits i medvetandet enligt det fjärde kriteriet, likaså eller. Ordet barn* kunde lika gärna ha stavats børn* . Att såväl barn* som för hänförts till danska kan bero på att den IBM-stans som använts inte har det danska ø-tecknet. Varje referens som innehållit detta tecken har stansoperatrisen förändrat till ett ö eller det svenska ordet. Det finns troligen fler fall än vad som registrerats via vår sökteknik.

Tabell 3. Språkbestämning av referenser:
Resultat från sex språk. Hela materialet

	Eng f	%	Sv f	%	Ty f	%	Fr f	%	No f	%	Da f	%	Övr f	%	Σ f	%
Korrekta utfall	5911	50	3341	28	666	6	145	1	72	.6	42	.4			10177	86
Felaktiga utfall	4	0	44	0	12	0	16	0	4	0	1	0			81	1
Övriga													1573	13		13
Σ	5915	50	3385	28	678	6	161	1	76	.6	43	.4	1573		11831	

Procenttalen är beräknade på totalsumman = 11 831

Ruta 9. Felprotokoll

Automat. klassif.	Antal	Korrekt klassif.	Orsak (sökord)	Post nr	Orsak (annan)	Kommentar
1. Engelsk	2	dansk	of	(2)	felstans	og —— of
	1	svensk	study	(2)		pilot-study
	1	tysk	and	(2)	felstans	and —— und
2. Svensk	11	norsk	*psykologi*	(2)		väntat
	9	norsk	*pedagogik*	(2)		väntat
	4	dansk	*psykologi*	(2)		väntat
	3	norsk	och	(6)		ej väntat
	2	norsk	och	(2)		ej väntat
	2	dansk	*pedagogik*	(2)		väntat
	2	norsk	hos	(2)		medveten om
	2	dansk	barn*	(2)		medveten om
	2	tysk	*psykologi*	(2)		psykologie
	1	dansk	eller	(2)		medveten om
	1	dansk	för	(2)	felstans	av "förr"
	1	norsk	* svensk*	(2)		ej väntat
	1	norsk	*undersökning*	(2)		ej väntat, lapsus
	1	tysk	*års*	(6)		i årsbok
3. Tysk	10	latin	Berlin, Leipzig	(3)		tillhör "övriga" kort 3 går ej
	1	finsk	zur	(2)		tysk övers av titeln inom / /
	1	norsk	ein*	(2)		blev Einar
4. Fransk	11	latin	Paris	(3)		ej väntat
	2	svensk	dans	(2)		medveten om men ej antaget
	2	engelsk	Paris	(3)		medv (OECD)
	1	italiensk	l'	(2)		väntat
5. Norsk	2	svensk	Oslo	(6)		finländsk förf
	1	engelsk	Oslo	(3)		
	1	svensk	Oslo	(3)		utgivn
6. Dansk	1	engelsk	Köpenhamn	(3)		svensk förf

En författare har, som vi vet, läst latin, vilket vi inte har sökord för. Det är främst franska och tyska ortnamn som döljer de latinska verken.

De resterande felen är av sådant slag att det är svårt att gardera sig mot dem, t ex omedvetna felstansningar eller felstavningar i ursprungsmaterialet. Dessutom är flera förekommande språk så fåtaliga till representationen att det inte är rimligt att försöka gardera sig mot eventuella förekomster. Att det svenska substantivet 'dans' skulle förekomma antogs inte (kriterium 3). Misstaget blev helt klart när vi upptäckte referensen i fråga: "Ingen dans på rosor" som ju tillhör den senare tidens populär-psykologiska romaner.

2.6 De specifika posternas andel i språkbestämningen

Vi har nu gjort våra explorationer i materialet. Vi började med en uppsättning sökord från de poster som anger tryckort (3), förlag (4) och tidskrift/institution (6). Vartefter försöken fortskred såg vi att ortsangivelser (oavsett post) inte ger entydiga utfall, vilket till sist medförde en ändring av sökrutinen till att gälla titelposten (2) i första hand. Det lärde oss bl a att vi inte behöver ha så stort lexikon att matcha texten mot vid bestämningen. Vi har också sett att det går att få fram en kärna av språkspecifika sökord, som tillsammans med andra mera ändamålsenligt antagna fungerar för detta bestämda syfte. Dessutom har de olika posterna kompletterat varandra. Där titelposten inte kunnat användas för bestämning har övriga poster tagits till hjälp. Detta kapitels slutredovisning innebär att vi ska titta på hur posternas användning har fördelat sig i vår stora sökning. Enligt utfallet som redovisades under kapitel 2.4 har post 4 utgått. Den totala summan i tabell 4 anger antalet korrekta utfall (jfr tabell 3, där både fel- och nollutfall redovisats).

Tabellen visar att 96 % av de korrekta utfallen tas ut via titlarna. I endast 4 % av fallen behöver vi ta hjälp av referensspecifika "cues".

Vad vi dessutom kan läsa ut ur tabell 4 är att det råder en intressant skillnad mellan dels nordiska och övriga referenser, dels svenska och övriga. När vi ska bestämma tyska, engelska och franska referenser tar vi (i ordning) post 3 respektive post 6 till hjälp, ifall post 2 inte ger utslag. I fråga om svenska referenser går det inte att använda post 3 (sökord finns inte), eftersom t ex Stockholm resulterar i såväl svenska

Tabell 4. Posternas betydelse för utfallet av språkbestämning av referenser

	Antal korrekta utfall för språken													
	Eng		Sv		Ty		Fr		No		Da		Σ	
	f	%	f	%	f	%	f	%	f	%	f	%	f	%
Post 2 (titel)	5733	56	3136	31	639	6	113	1	7	0	2	0	9630	96
Post 3 (ort)	163	2	-	-	22	0	21	0	56	0	35	0	297	2
Post 6 (inst)	15	0	205	2	5	0	11	0	9	0	5	0	250	2
Σ	5911	58	3341	33	666	6	145	1	72	0	42	0	10177	100

Procenttalen är beräknade på totalsumman = 10 177 (korrekta, tab 3)

som engelska böcker (avhandlingar, läroböcker i översättning etc). Den svenska siffran under post 6 ska här tydas som att det refereras en hel del icke tryckt material eller utredningar. Tidskrifter kan vara fackpress, årsböcker etc (se ruta 7). Problemet med att de svenska institutionsserierna avfattas på annat språk än vad serienamnet anger (t ex Lund) har däremot försvunnit genom att vi orienterade oss mot post 2 istället.

En annan intressant skillnad är att norska och danska referenser bäst tas ut via post 3 och därnäst post 6. Det är alltså tryckorterna (Oslo, Köpenhamn) som tar upp nästan allt för dessa språk. Det har bl a den fördelen att vi kan undvika problemet med att specificera sökord för att skilja norska från danska, som är mycket lika i sina formord.

Vi kan också konstatera att de engelska referenserna lättast kan bestämmas via sökord på titlar. Därnäst följer svenska och tyska. Ju fler ord (eller tecken i sekvens) ett språk har gemensamt med ett annat språk (t ex grammatiska likheter i de nordiska språken eller i italienskan och franskan), desto svårare blir det att automatiskt bestämma språket i material av det här slaget. Men sett ur ett internationellt perspektiv kanske en databas med t ex nordiska referenser mycket naturligt skulle tillhöra samma kategori.

3. LEXIKON FÖR SPRÅKSEPARERING

Vi har nu efter flera utprövningstillfällen kunnat utkristallisera vissa sökord, som bäst separerar referenser på olika språk från varandra. Sökorden finns i olika s k poster inom en referens, beroende på betydelse i sammanhanget. Vissa sökord är t ex till för att bestämma dokumentets tryckort, andra bestämmer om det är fråga om tidskriftsartikel, institutionspublikation etc. Dessa kan kallas referensspecifika sökord, dvs sådana som finns därför att refereringstekniken bjuder så. Sedan har vi sådana som är dokumentspecifika, då främst titeln. Inom titeln finns språkspecifika sökord, som säger på vilket språk dokumentet är skrivet, samt ämnesspecifika sökord, som anger området. (Grafotaxen har vi inte beaktat här.)

I föregående kapitel visade tabell 4 de olika posternas betydelse för bästa utfall vid separeringen. Det här kapitlet ska tala om vilka sökord inom respektive post som bäst kan användas, dvs som "tar upp" de flesta referenser i ett beteendevetenskapligt dokumentmaterial av intresse för utbildningssektorn. Här påminns läsaren om sökteknikens betydelse för frekvenserna, vilket är viktigt när det gäller titelposten: Först påträffade sökord ger utslag, dvs referensen sorteras in i respektive språkbas (får en kod) och respektive sökord får en frekvensmarkering.

Presentationen avser frekvenserna i tabell 4, korrekta utfall.

3.1 Referensspecifika sökord

De sökord som är specifika för referenserna och som haft betydelse här finner vi inom posterna 3 och 6, som anger tryckort respektive tidskrift/institution, etc.

Tabell 5. Frekvenslista för sökord inom post 3 (tryckort)

Sökord	f	%	Sökord	f	%
New York	101	34	Berlin	5	2
Oslo	51	17	Kristiania	4	1
London	48	16	Oxford	3	1
Paris	21	7	Odense	2	1
Köpenhamn	20	7	Frankf. *	1	0
Leipzig	16	5	Krist. *	1	0
Köbenhavn	13	4	Pittsburgh	1	0
Chicago	9	3	San Francisco	1	0

* Förkortningarna förekom i materialet.
Procenttalen är beräknade på totalsumman för post 3 (tab 4) = 297

Här framgår att New York och London tar nästan allt som publicerats på engelska. Oslo är, som vi kunde vänta oss efter en blick i tabell 4, av stor betydelse för norska dokument. För de övriga språken har post 3 mindre betydelse. Svenska dokument kommer inte ut förrän vi söker på post 6. Detta framgår också ur tabell 6:

Tabell 6. Frekvenslista för sökord inom post 6 (tidskrift, etc)

Sökord	f	%	Sökord	f	%
svensk	57	23	norsk*	5	2
och	40	16	Oslo	4	2
års*	29	12	schrift	4	2
SOU	28	11	dansk*	3	1
lärar*	27	10	Kbhn	2	1
review	15	6	prop*	2	1
utredning*	12	5	revue	2	1
Paris	9	4	und	1	0
arbet*	6	2			

Procenttalen är beräknade på totalsumman för post 6 (tab 4) = 250

Här ser vi att de svenska sökorden dominerar, vilket troligen beror på att denna post till stor del specificerar "icke tryckt" material, utredningar m m. Ett formord, och, går bra här. I övrigt lägger man märke till sökorden *svensk* , års* , SOU och lärar* , som tydligt visar vilket material det är fråga om. De övriga språken har mycket liten representation inom post 6. Vi ser också att engelska språket endast representeras av ett enda sökord inom denna post. Det betyder att typen av engelska dokument är enhetligare än t ex de svenska (se tab 4).

3.2 Språkspecifika sökord

Vi börjar presentationen av sökorden för titlar (post 2) med det högst frekventa språket och går nedåt. Observera att titelkortet tar upp 96 % av alla korrekta utfall (se tab 4).

Tabell 7. Frekvenslista för engelska sökord: post 2 (titlar)

Sökord	f	%	Sökord	f	%	Sökord	f	%
the	1551	17	measur*	93	2	how	25	0
of	1028	18	child*	92	2	into	24	0
and	902	16	learning	86	1	adult*	23	0
study	239	4	some	83	1	its	17	0
on	237	4	psychological	76	1	training	17	0
psychology	147	3	studies	67	1	by	15	0
education	146	2	methods	65	1	America*	13	0
analysis	132	2	as	56	1	is	13	0
educational	116	2	personality	44	1	or	12	0
teach*	104	2	from	34	1	towards	10	0
school*	103	2	with	32	1	who	7	0
research	96	2	reading*	28	0			

Procenttalen är beräknade på totalsumman för post 2 (tab 4) = 5 733, engelska

Här dominerar tre språkspecifika formord, nämligen the, of och and, som tar upp drygt 60 % av det engelskspråkiga materialet. De sökord som därnäst är användbara är de ämnesspecifika. psychology, education och school* t ex är mycket vida begrepp, som anger en hel sektor. Ur informationssynpunkt ger de därför mycket lite ur en dokumentbas för utbildningsforskning.

Att formorden har få "träffar" relativt sett beror bl a på att inte alla ord kan förekomma i titelns början.

Tabell 8. Frekvenslista för svenska sökord: post 2 (titlar)

Sökord	f	%	Sökord	f	%	Sökord	f	%
och	757	24	från	59	2	vux*	26	1
skola	251	8	läro*	57	2	beteende	25	1
svensk	217	7	*historia*	56	2	eller	23	1
för	209	7	att	52	2	hos	19	1
psykologi	179	6	Sverige	52	2	inom	21	1
till	165	5	*pedagogik*	45	1	*skolor*	21	1
barn*	116	4	*mätning*	31	1	personlighet	19	1
undersökning	96	3	betänk.	29	1	hur	18	1
utbildning	87	3	mellan	29	1	*pedagogiska*	16	0
års*	86	3	vad	29	1	*utredning*	15	0
ett	80	3	uppfostr*	27	1	är	13	0
arbet	75	2	ur	27	1	samt	7	0
någ*	71	2	begåvning*	26	1	vem	7	0
						inför	4	0

Procenttalen är beräknade på totalsumman för post 2 (tab 4) = 3 136, svenska

Vi ska vara tacksamma för vårt och. Det är ett språkspecifikt ord som har mycket stor separeringsförmåga. Flera steg efter kommer några ämnesspecifika, nämligen *skola*, *svensk*, *psykologi* och barn*, som sammanfattar området, precis som i den engelska listan. Ett par prepositioner befinner sig bland dem, nämligen för och till. En hel del formord har inte heller i svenskan särskilt hög frekvens sett ur vårt söktekniska perspektiv. *pedagogik* förekommer mera sällan i utbildningsforskarnas referenslistor än *psykologi*. Vad ett sådant resultat återspeglar kan vi här inte säga. I så fall krävs en annan slags frekvensräkningsteknik än den sorteringsmekanism som används här.

Tabell 9. Frekvenslista för tyska sökord: post 2 (titlar)

Sökord	f	%	Sökord	f	%	Sökord	f	%
die	108	17	werk*	13	2	pädagogik	7	1
und	84	13	untersuchung*	11	2	pädagogisch	7	1
über	71	11	bei	10	2	oder	6	1
zur	57	9	schul*	10	2	schrift*	5	1
das	42	6	vom	10	2	unterricht*	5	1
geschichte	35	5	deutsch*	9	1	nach	3	0
ein*	28	4	für	9	1	*forschung*	2	0
erziehung*	23	4	psychologisch*	9	1	mit	2	0
buch	17	3	als	8	1	ohne	2	0
bildung*	14	2	im	8	1	zum	2	0
jugend	14	2	*wissenschaft*	8	1	auf	1	0

Procenttalen är beräknade på totalsumman för post 2 (tab 4) = 639, tyska

Fem formord toppar den tyska listan, bl a beroende på det tyska språkets "explicithet". Dessutom får det historiska större betydelse här genom Geschichte. Ämnesområdet är äldre på tyskt språkområde än på svenskt. Schul* har relativt sett lite mindre betydelse för tyskt språk än för svenskt och engelskt, men det är ändå ungefär samma ämnessfär som kommer till uttryck, t ex Erziehung* och Jugend*.

Danska och norska har endast varsitt sökord inom titeln och ingår ej i någon tabell. (Orden var "norsk" och "undersøgelse" med 7 respektive 2 frekvenser.)

Det återstår endast några få frekvenser för de franska sökorden.

Tabell 10. Frekvenslista för franska sökord: post 2 (titlar)

Sökord	f	%	Sökord	f	%
l'	40	35	ou	5	4
sur	18	16	française	3	2
les	12	10	pour	2	2
oeuvre*	10	9	un	1	1
chez	10	9	une	1	1
France	6	5			
dans	5	4			

Procenttalen är beräknade på totalsumman för post 2 (tab 4) = 113, franska

De franska sökorden är få, eftersom vi inte har velat få en sammanblandning med engelska eller italienska. Strängen *ique* kunde t ex inte användas. l' har dock så stor effekt att det ändå fått kvarstå.

Bestämda artikeln l' dominerar här, liksom i engelskan, följd av sur och les. Det man läser på franska är oeuvre* (jfr ty Werk*) men det går inte fram vad de handlas om. Några ämnesord finns inte alls här, p g a för stor sammanblandning med andra språk (jfr 'psychologie'). Användningen av diakritiska tecken har inte tillåtits i vår utrustning (jfr e'ducation-education). Tecknet ç (française i tab 10) har alltså stansats c och accenttecknet i egen kolumn (apostrof).

3.3 Olika söknivåer

I kapitel 2 angavs i samband med att kriterier uppställdes för ett språksepareringslexikon att "diskriminatorer" kan finnas på en tredje nivå, nämligen den grafotaktiska. Det finns studier som redovisar hur grafem (ung. bokstav) kan inta unika kombinationer, dvs grafemens syntaktiska regler hindrar eller befrämjar vissa sammanställningar. Sådana kombinationsstudier av autografemens (vokal-) och syngrafemens (konsonantkombinationer) strukturer kunde göras i och med att datorer kom till användning i språkforskningen.

En teckensekvens, som vi använder ordet, kan bestå av digram (tvåkombinationer), trigram (trekombinationer), etc. Dessutom kan man specificera var i sekvensen kombinationen ska förekomma; i början, i slutet, som första digram efter första vokal, etc. Frekvenser för

sådana kombinationer redovisas och diskuteras i Allén (1971, kap 5).

Datamaskinens användning har också gjort det möjligt att, baserat på en modell om fonem, stavelser eller morfem, "tillverka" nya svenska ord som är lediga och som uppfyller de krav som språkmodellen för svenska stavelser etc ställer. Detta har utnyttjats i reklamen vid bildandet av varunamn (Sigurd, 1970).

Vad hade det inneburit för vår analys att hålla oss på den grafotaktiska nivån? Ja, vi hade troligtvis inte kunnat göra just den här typen av jämförande analys utan ett omfattande merarbete i att försöka anskaffa grafotaktiska uppgifter för olika språk, som inte baseras på frekvenser. Att utveckla program för att utkristallisera unika grafemkombinationer hade inte varit rimligt. Ämnesorden gav i denna studie en god uppfattning om området, vilket en konsonantkombination självfallet inte gör. Dessutom baseras vår sökteknik på sökning "från vänster". Ett letande efter en unik kombination av auto- och eller syngrafem hade troligen tagit längre tid och blir dyrt när det gäller så här stora textmängder. Däremot hade det varit intressant i ett projekt av annan karaktär.

4. NÅGRA KONTROLLER OCH JÄMFÖRELSER

4.1 Verk och referenser

Separeringsprogrammet, såsom det redovisades i ruta 8, har testats på verkbasen. För en jämförelse med referenserna (hela materialet i tabell 3, s 23) presenteras en översiktstabell över utfallet på hela verkbasen.

Tabell 11. Språkbestämning av verk: Hela materialet

	f	%
Korrekta utfall	716	90.18
Felaktiga utfall	3	.38
Övriga	75	9.45
Σ	794	

Tabell 11 visar att drygt 90 % av utfallen blir korrekta. Det är en ökning med 4 % jämfört med referensbasen (tabell 3). De verk som inte bestämts via programmet är därmed också färre än i referenserna. Även andelen fel minskar. Orsakerna ligger närmast i att forskarna inte producerar sig på lika många språk som de läser (eller refererar). Basen innehåller främst svenska och engelska i nu nämnd ordning. Sökorden för övriga språk har inte i nämnvärd grad kunnat fyllas upp med frekvenser.

En jämförelse mellan de två baserna ifråga om svenska och engelska sökord har kunnat göras. Följande kunde noteras: Det råder stor överensstämmelse mellan proportionerna per sökord mellan de båda baserna. Men några få skillnader kan synas intressanta.

I forskarnas egna egna titlar tar sökorden för, *pedagogik* och *mätning* upp fler dokument än i referenserna. (Skillnaden är då minst tre procentenheter.) I dessa tar däremot *svensk* och barn* större andel.

När forskarna producerar engelska titlar blir det främst educational

och study som skiljer sig från referenstitlarna. Det engelska materialet tas bättre upp av the och psychology. I verktitlarna har ingen frekvens noterats för psychology eller *historia, som båda finns i referenserna.

Vad dessa få skillnader betyder är svårt att säga f n. Att procenttalen ändrar sig när det gäller några få sökord torde främst ha att göra med dispositionen av titeln (ord i början får frekvenser). Men först efter ingående studium kan vi säga om eventuella skillnader avspeglar innehållsliga eller strukturella förändringar, t ex genom att ta in en tidsdimension.

Kontrollen av programmet har givit det resultatet att vi kan använda samma sökord på verk och referenser och räkna med ett likartat utfall. Dessutom är variationen inte så stor i verktitlarna som i referenstitlarna, vilket gör att en sökning i verken ger lite bättre utfall.

4.2 Manuell och automatisk språkbestämning

Vid kodningen av de 40 forskarnas verk utnyttjades en sjunde post för dokumentkarakteristika av annat slag än de bibliografiska uppgifterna (se ruta 1, s 4). Inom post 7 kodades bl a om dokumentet är svenskt eller översättning till annat språk. Kodbeteckningen är svensk = 1, annat språk = 2. Upplysning om vilket det andra språket är ansåg vi oss inte behöva. Vi har ju också sett i kapitel 4.1 att det främst är engelska som översättningen avser (författarna är svenska).

Vid kodningen har en person använt fotostatkopior av titelbladet/första sidan av verket. Det är inte troligt att någon annan uppgift än själva titeln kommit till användning vid språkkodningen.

En jämförelse mellan den manuella kodningen, omfattande 2 koder (typ "ja-nej") och den automatiska, omfattande 6 koder har gjorts maskinellt och resulterade i en korstabell, som visar andelen överensstämmande kodning. Om man går in i tabellen i raden för "svensk" och läser av kolumnen för "svensk" så avläses hur stor överensstämmelsen är. Sammanlagt skiljer sig den manuella och den automatiska kodningen till 2.8 %. På vilket sätt visas i tabell 12.

Tabell 12. Jämförelse mellan manuell och automatisk språkkodning av verk: Felanalys

	Felkodning							
Kodning	Sv	No	Da	Ty	Eng	Fr	Σf	%
Automatisk	-	-	1	1	1	-	3	.4
Manuell	4	-	-	2	11	-	17	2.4
Σ	4	-	1	3	12	-	20	2.8

Totala antalet verk = 794

Antalet fel följer inte språkens procentuella andel i verkbasen: De engelska felen är betydligt fler än de svenska. Dessutom finns de svenska felen inte i den automatiska kodningen. I den manuella kodningen har 11 engelska titlar kodats som svenska. Det händer inte med det automatiska programmet.

Det automatiska engelska felet är intressant. Samma fel har nämligen kodaren också gjort. Titeln lyder: "The cycling strength test (CST) som prediktionsinstrument vid prövning av skyttesoldater". När vi byggde upp det automatiska programmet försökte vi förhindra sådana fel genom att införa en hierarkisk ordning i sökproceduren (se ss 16 och 19). Detta har tills vidare inte gällt, vilket ger till resultat att the blir nyckelord för insortering i den engelska basen. Vår kodare tycks ha gått tillväga på samma sätt, åtminstone har inte fler än de fyra första orden lästs.

De övriga felen studeras inte närmare här. Vid ett ytligt betraktande är det svårt att förklara de manuella felen. En trötthetseffekt kan ha inträffat. Det är monotont att bara hålla reda på koderna 1 eller 2. Vid flera koder är det kanske lättare att behålla uppmärksamheten? Stansningen är kontrollerad, vilket utesluter överföringsfel. (Stansning utförs av institutionens stansoperatris, som inte är samma person som kodaren.) En stanskontroll är rapporterad inom det anslutande projektet, där det bl a konstaterades att de flesta fel görs vid text, trots att de numeriska koderna är fler (I. Bierschenk, 1974, s 33).

Andelen fel (2.8 %) bör betraktas som liten. Men inom den felmarginalen (som kan öka vid manuell kodning med större material) bör skillnaden (2 %) mellan manuell och automatisk kodning uppmärksammas.

F n har vi i jämförelsen uteslutit den del där vi ännu inte har utvecklat rutiner, t ex subrutiner i sökningen med hjälp av ett antal villkor (9.4 % av materialet i verkbasen). Automatisk kodning tycks trots detta vara överlägsen, vilket betyder att en utveckling av dessa rutiner kan och bör fortsätta.

4.3 Äldre och nyare data

En kompletterande datainsamling har skett sedan dokumentbaserna lades upp. Denna insamling gjordes för att täcka de senare årens produktion (1975-77) hos de 40 forskarna. Eftersom I&D-projektet ansluter sig till ett tidigare projekt, där dessa forskare ingår kom datainsamlingen att avslutas med detta projekts årtalsgräns (1974). Men det torde vara värdefullt att det nya projektets data är så aktuella som möjligt. Dessutom har det talats, åtminstone i vissa grupperingar, om ett s k "paradigmbyte" i forskningen omkring 1974, och om så är fallet finns det ju anledning att studera om detta återspeglas i litteraturen.

Separeringsprogrammet har prövats på denna nytillkomna referensbas. Storleksordningen mellan språken är densamma som tabell 3 (s 23) visar, även om andelen svenska referenser tycks ha ökat. För en jämförelse av programmets verkan på den äldre och den nyare basen görs en sammanställning i tabell 13.

Tabell 13. Kontroll av språkseparering: Äldre och nyare referenser

	Referensbaser (1937-74) %	(1975-77) %
Korrekt	86	87.5
Fel	1	.5
Övr	13	12
Σ	100	100

Bastal n = 11 831 respektive 4 062

Som vi ser i tabell 13 går det knappast att tala om några skillnader. Om skillnaden som trots allt kan utläsas verkligen återspeglar en förbättring, så bör den väl tolkas så att en del litteratur som orsakar besvär för programmet (för det mesta äldre litteratur) inte har citerats under denna period.

Studier i lexikonen visar på att några ämnesords separeringsförmåga (frekvens) ändrar sig. Det är fortfarande samma ord som är gångbara, men de byter rang. T ex ökar de engelska educational, education och research. För tyskan märks ingen speciell ökning av något ord. Istället minskar Geschichte och Werk* i betydelse. I de svenska referenserna märks främst en ökning i *skola*, *utbildning* och *utredning*. Om förändringar av det här slaget verkligen är förändringar av referensramar kan bl a vårt nästa resultat hjälpa till att belysa.

Det visar sig nämligen att formorden inte förändrar sig. Eftersom sökproceduren har gått till på samma sätt som tidigare och strukturen i titlarna tycks vara konstant, kan studier av förändringar i begrepp bli meningsfulla.

5. SAMMANFATTNING

Med ledning av de explorationer och kontroller som gjorts ska detta kapitel sammanställa några punkter som vägvisare för det fortsatta arbetet med att bygga upp ett informationssystem inom utbildningssektorn i Sverige.

1. Språkseparering

Det material som är tänkt att tas in och bearbetas kommer att behöva underkastas en hel del automatiska förfaranden. Vid stora mängder referenser är det av stort värde att separera dokument i fråga om språk. Våra studier har visat att redan ett enkelt program fungerar bättre än en manuell kodning. Ett sådant rutinarbete bör läggas över på en dator. Människan är helt enkelt inte reliabel. Man kan också uttrycka det med att maskinen inte kan fantisera medan den utför instruktioner.

Att utveckla klassificeringsrutinerna på detta område är väsentligt och har ett generellt värde, inte minst med tanke på att många informationssökningssystem som utvecklats inom olika språkområden ska kunna sammankopplas utan att för mycket brus uppstår i baserna.

Förutsatt att kontroller av separeringsprogrammet görs med jämna mellanrum har vi genom dessa automatiska rutiner möjlighet att upptäcka när förändringar inträffar i sättet att skriva referenser, att strukturera titlar, att tackla ett problemområde eller förändringar i publiceringspolicyn. Det skulle vi knappast märka om vi kodade manuellt.

2. Experimentbaser

För ett utvecklingsarbete av det här slaget krävs ett omfattande experimenterande som rör språkvetenskapliga, datatekniska, statistiska och i vårt fall beteendevetenskapliga frågeställningar. Det är ofta praktiskt att ha ett mindre, hanterbart material för detta ändamål. Våra studier har visat att verkbasen inte skiljer sig nämnvärt från den stora referensbasen ifråga om sökord i titlar. Studier på nyinsamlade data visar dessutom att det inte råder någon skillnad mellan den stora referensbasen och den nyinlagda ifråga om programmets precision. Några större skillnader i procenttalen mellan språken finns inte heller.

Det gör att olika resultat från experiment i de mindre databaserna är möjliga att generalisera till den större basen.

3. Strukturer i vetenskapliga dokumenttitlar

Den sökteknik vi utprövat för snabb språkseparering resulterar i frekvenser av sökord som har samband med de strukturer som en titel har. Vissa formord är t ex vanliga i början, andra längre fram. Man kan förmodligen också tala om givna mönster: "Om titeln innehåller tre formord och det första är x, så är sannolikheten si eller så stor att de andra två är y och z i nu nämnd ordning", etc. Bland dessa högfrekventa s k språkspecifika sökord kan vi urskilja ord av olika slag. Förutom formord har vi ett slags neutrala presentationsord av typen Studier av ..., En analys av ..., Forskning kring ... I detta material skulle de mycket väl kunna räknas till formorden. De tillhör sättet att presentera ett vetenskapligt arbete. Den tredje gruppen har kallats ämnesord som kan exemplifieras med psykologi, skola och education. Ur informationssynpunkt är ämnesorden ointressanta. Inte förrän de bryts ner i avgränsade delområden är de av intresse, t ex arbetspsykologi, förskola, special education. Men använda som sökord i en omfattande databas är risken stor att de tar upp mängder av dokument, eftersom de representerar vida sektorer inom utbildningen. Det är ju för differentieringens skull man vill skapa en lämplig sökprofil.

Vad en fortsättning bör kunna ge svar på ifråga om titlars informationsvärde är

(1) Hur är en titel strukturerad?
(2) Var i strukturen finns de informativa begreppen och hur är de uppbyggda? samt
(3) Hur ska dessa kunskaper presenteras i en tesaur för att så många som möjligt ska kunna använda den?

4. Förändringar över tid

Den sista kontrollen avsåg en jämförelse mellan den stora databasen av referenser från åren 1937-74 och en nyinlagd bas med en komplettering för åren 1975-77.

Det visade sig att formorden, som anger strukturen i titlarna, inte förändrar sig proportionellt från äldre till nyare data. Däremot tycks vissa ämnesord ändra sig. Detta har stora implikationer för fortsatta

studier av begrepp och begreppsrelationers förändringar. Om vissa begrepp används ofta eller på nytt sätt på senare år, så finns det anledning att förmoda att sådana förändringar kan spåras även i den stora referensbasen. Ett intressant angrepp kan vara treårs-cyklar för sådana trendanalyser. Andra studier i materialet har nämligen visat att treårsintervall är en meningsfull gruppering. Det är t ex den vanligaste anslagsperioden i projektforskningen, vilket märks i rapporteringen.

6. REFERENSER

Allén, S. Introduktion i grafonomi. Stockholm: Almqvist & Wiksell, 1971.

Allén, S., Järborg, J. & Ralph, B. Svensk ordbok och lexikalisk databas. Förstudierapport. Stencil (Göteborg: Inst. för språkvetenskaplig databehandling), 1977.

Bierschenk, B. Handledning för rapportering av beteendevetenskaplig forskning. Pedagogisk dokumentation, Nr 18, 1973.

Bierschenk, B. Perception, strukturering och precisering av pedagogiska och psykologiska forskningsproblem på pedagogiska institutioner i Sverige. Pedagogisk-psykologiska problem, Nr 254, 1974.

Bierschenk, I. Konstruktion av ett regelsystem för en datorbaserad innehållsanalys av intervjutext: Preliminärmanual och några utprövningsresultat. Testkonstruktion och testdata, Nr 25, 1974.

Sigurd, B. Att tillverka ord. I: Allén, S. & Thavenius, J. (Red.) Språklig databehandling. Lund: Studentlitteratur, 1971. Ss 77-85.